**总复习**

**【实验目的】**

**复习所学HDFS，Rdd和DataFrame的内容**

**【实验内容】**

**一、HDFS的操作：**

**1、在hdfs中新建目录test，并通过命令查看创建结果；**

**2、将本地盘/root/test下方的test1.txt上传到hdfs的myspark目录下方，并查看结果；**

**4、在hdfs上查看test1.txt文件的内容，通过命令查看结果；**

**5、将hdfs上myspark目录中的test1.txt下载到虚拟机本地盘/usr/local/下方，并查看结果；**

**6、删除hdfs上的test1.txt文件。**

**7、删除hdfs上创建的目录myspark，通过命令查看结果。**

**二、Rdd的创建方法**

**1、将本地磁盘/root/test下方的student.txt创建成rdd1，并显示查看。**

**2、将文件student.txt上传到HDFS的myspark目录下，使用相应方法将其转换成rdd2，并显示查看；**

**3、将10以内的奇数生成列表并转换为rdd3，并显示查看。**

**三、spark之RDD的创建与转换操作。**

**1、将本地磁盘/root/test下的word.txt转换成rdd10，并打印输出其结果后保存为write文件；**

**2、将文件word.txt转换的rdd10使用map转换，使用空格分离转换，并查看其结果。**

**3、将文件word.txt转换的rdd10使用flatMap转换，使用空格分离转换，并查看其结果。**

**4、使用groupByKey()转换对rdd10进行转换，并查看结果；**

**5、使用reduceByKey()转换对rdd10进行求词频，并查看结果；**

**6、将文件student.txt转换成RDD，并使用持久化对行动操作count(),first(),collect(),take(n) 查看结果。**

**6、将test1.txt文档转换成rdd，并设置为3个分区，命名为rdd；**

**7、显示分区个数；重新设置分区个数为2，并显示。**

**8、在rdd1中筛选出包含“spark”的字符串保存到新的rdd2中，并打印显示；**

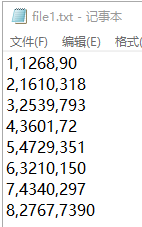
**四、DataFrame创建**

**1、将本地路径/root/test下的文件people.json，people.txt，11.csv,分别使用2种方式创建DataFrame，并显示输出；**

**2、将1题已经创建好的DataFrame保存成newpeople.json，newpeople.txt，并查看保存成功与否。**

**五、Rdd应用题**

**在/root/test目录下有2个文本文件file1.txt和file2.txt，每个文件中有很多行数据，每行数据由3个字段的值构成，不同字段之间用逗号隔开，如下图所示。其中3个字段分别是：orderid，userid，payment，现要求通过rdd的各种转换操作实现求出payment字段Top值的前5个。**



**六、**

**/root/test下现有学生信息文档student.txt，课程信息文档course.txt和成绩信息文档grade.txt。**

**1、查询181005学号学生的姓名和年龄；**

**2、查询孙慧选修课程的课号和成绩；（嵌套查询）**

**3、查询选修大学物理课程的学分、课号和成绩；（连接查询）**

**4、查询选修离散数学课程的学号、姓名、课号。（内连接查询）**